Reference-free high-throughput SNP detection in pea: an example of discoSnp usage for a non-model complex genome

Susete Alves Carvalho, Raluca Uricaru, Jorge Duarte, Claire Lemaitre, Nathalie Rivière, Gilles Boutet, Alain Baranger, Pierre Peterlongo

Reference-free high-throughput SNP detection in pea: an example of discoSnp usage for a non-model complex genome

Susete Alves Carvalho[1], Raluca Uricaru[1,3,4], Jorge Duarte[2], Claire Lemaitre[3], Nathalie Rivière[2], Gilles Boutet[1], Alain Baranger[1] and Pierre Peterlongo[3].

[1]UMR1349 IGEPP INRA, AGROcampus-Ouest Université Rennes 1, 35653 LE RHEU France
[2]BIOGEMMA, Upstream Genomics Team, 63720 CHAPPES France
[3]INRIA Rennes, Bretagne Atlantique/IRISA, EPI GenScale, 35000 Rennes, France
[4]LaBRI, University of Bordeaux , 33400 Talence, France

Introduction

Detecting Single Nucleotide Polymorphisms (SNPs) between genomes is a routine task with Next Generation Sequencers (NGS) data. SNP detection methods generally need a reference genome. As non-model organisms are increasingly investigated, reference-free methods are needed. The **discoSnp method** (Uricaru *et al*.) detects SNPs directly from raw NGS data set(s) without using any third-party information.

The **pea non-model organism** has a 4.5 GB complex genome without reference. We compared, on **the same set of low depth pea sequences**, the SNPs generated by **discoSnp** with those published with a **previous SNP discovery pipeline** (**Duarte *et al***. 2014), and those generated using **classical mapping approach** with the association of **Bowtie2 and GATK** tools.

SNPs data by *Duarte et al. 2014*

35,455 SNPs defined *in-silico* from normalized cDNA sequencing (454 GF-FLX sequencer) of 8 pea genotypes.

From them:

- ✓ **1,920 genotyped** with an *Illumina Golden Gate* assay and **84% succeeded**.
- ✓ **1,340 mapped** on a Pea consensus genetic map and their reference contigs anchored to the model species M.*truncatula* physical map.

discoSnp software

- ✓ Research SNPs from **any number of raw NGS dataset(s)**
- ✓ **No reference genome, data assembly or specific annotations needed**
- ✓ Composed by two modules to detect SNPs and improve their robustness.
- ✓ Availability: http://colibread.inria.fr/software/discosnp/

42,645 SNPs defined *in-silico*.

Bowtie2+GATK tools

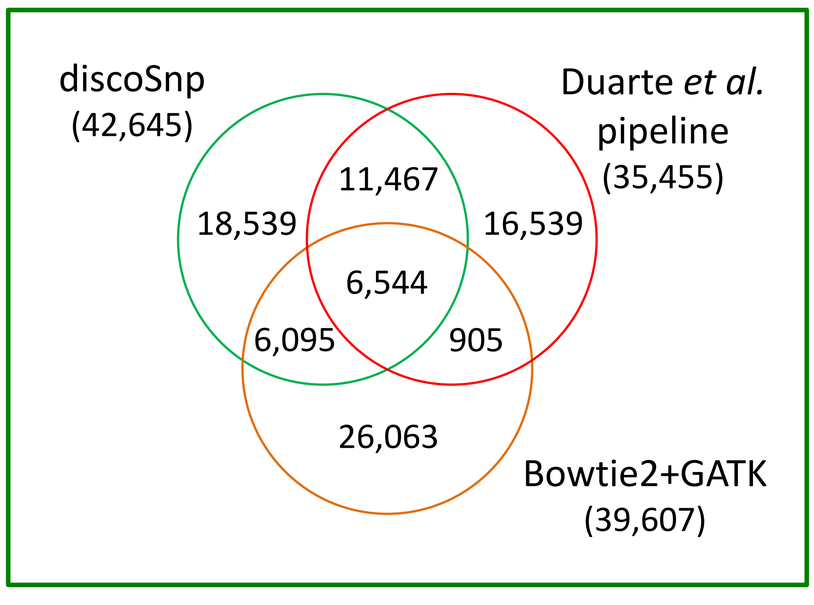
- ✓ Mapping reads in contigs developed by Duarte *et al.* 2014. (Bowtie2)
- ✓ SNP detection in mapped outfile. (GATK)
- ✓ Availability:

http://bowtie-bio.sourceforge.net/bowtie2

https://www.broadinstitute.org/gatk

39,607 SNPs defined *in-silico*.

Results



1) Comparison of 3 methods *in-silico*

- ✓ Large numbers of SNPs were identify by either methods
- ✓ **6,544 SNPs common to all data sets**
- ✓ **18k, 12k and 7k SNPs** were common to discoSnp vs Duarte *et al.*, discoSnp vs Bowtie2+GATK and Duarte et al. vs Bowtie2+GATK, respectively.

2) Comparison of the methods on SNPs *in-vivo* validated

- ✓ From 1,920 SNPs genotyped by Duarte *et al.*: **1,271** found by **discoSnp** and **590** found by **Bowtie2+GATK**.
- ✓ From 1,340 SNPs mapped on a Pea consensus genetic map: **929** found by **discoSnp** and **432** found by **Bowtie2+GATK**.

→ The three methods each provide between 35k and 40k SNPs but only 6k SNPs are common between the three sets. Between 16k and 26k SNPs belong to only one detection method specifically.

→ The reason why SNPs specifically generated by either method were lost in the genotyping process may be due to: **i/** filtering parameters including low coverage of the sequencing data, **ii/** the large "kmer" size (especially for **discoSnp, it does not to detect SNPs that are too close one from others**).

Conclusions and Prospects

- ✓ In conditions of unsequenced genomes and low sequencing coverage, different tools on the same data can generate complementary SNPs sets.
- ✓ discoSnp software offers the advantage to find robust SNPs without the need to perform assembly. It can therefore be successfully applied to non-model unsequenced genomes (Quillery *et al.* 2014).
- ✓ The quality of discoSnp results in association with its very low memory needs and low time footprints led us to choose this software for a SNP discovery and direct **Genotyping By Sequencing project** on a set of 48 pea genomic DNA libraries from a recombinant inbred lines subpopulation sequenced with Illumina HiSeq2000 technology. The analysis enabled to identify **88,851 SNP polymorphs** on this population, from which around **60k SNPs will be genetically mapped** (Boutet *et al.* 2014).

ECCB 14 STRASBOURG · FRANCE 7-10 SEPTEMBER 2014

Uricaru et al. (2014). Submitted
Duarte et al. (2014). BMC Genomics 15:126
Quillery et al. (2014). Molecular Ecology Ressources 14(2)
Boutet et al. (2014). IFLRC VI 2014; Canada; Oral Communication

The authors acknowledge the financial support of SOFIPROTEOL under the Project PEAPOL